

基于开放获取论文推送转发服务系统 iSwitch 的机构知识库内容建设*

张旺强　祝忠明　姚晓娜　刘　巍
(中国科学院兰州文献情报中心　兰州 730000)

摘要:【目的】将开放获取论文推送转发服务系统 iSwitch 分发的本机构知识产出数据自动同步存缴到机构知识库中。【方法】使用定时任务调度与 FTP 协议进行数据同步，通过文件包、文件解析将数据预加载到数据库，同时提供导入管理、已导入数据管理、审计等功能。【结果】实现数据的自动同步与半自动化导入。已完成对 Web of Science 超过 6 万条数据的接收与存缴。【局限】iSwitch 推送数据的准确率与及时性有待提高，IR 需进一步优化数据导入功能提高自动化程度。【结论】基于 iSwitch 的机构知识库内容建设，大大减轻了科研人员、机构知识库管理人员的负担并保证了数据质量。该模式具有一定的推广价值。

关键词: 开放获取　机构知识库　iSwitch　内容建设
分类号: G250.7

1　引　言

机构知识库(Institutional Repository, IR)的内容建设长期以来普遍面临着科研人员主动参与积极性不高的难题。根据欧洲 PEER 项目的调查报告，即使由出版商邀请作者上传论文的最终同行评议稿，实际存缴率也仅为 2%[1]。结合中国科学院 IR 的建设经验，科研人员之所以参与度不高，主要原因有科研人员不熟悉 IR 存储操作流程、担心付出的时间成本、IR 对用户的吸引力不足等[2]。

通过基于标准 Web 协议的机器接口(如 OAI、REST API、SWORD 等)从外部系统中获取本机构的知识产出逐渐成为一种比较流行的 IR 内容建设方式。这种方式可以减少资源重复建设带来的人力、物力消耗，也避免了人工存缴过程中可能造成的元数据丢失问题。

随着开放获取运动的发展，越来越多的出版商支持将论文的元数据(甚至全文)通过机器接口推送到作者所在机构的知识库中。例如，麻省理工学院图书馆与 BioMed Central 合作，后者会不定期自动将麻省理工学院作者发表的文章元数据与全文推送到他们的机构知识库中[3]；意大利国家地质与地球物理学研究所(INGV)[4]与开放获取期刊 *Annals of Geophysics* 达成协议，一旦前者有文章通过后者发表，后者会自动将论文提交到前者的 Earth-prints 知识库中[5]；JISC 的 Repository Junction Broker(RJB)项目[6]，旨在建立一个论文交换中心，它先从多个出版商系统中接收数据，再根据作者单位将论文分发到每个作者所在机构的知识库中。

中国科学院文献情报中心建立的论文推送转发服务系统——iSwitch[7]，其初衷类似于 JISC 的 RJB。主要功能是从相关出版社获取并按机构分发中国科学院

通讯作者：张旺强，ORCID：0000-0002-5105-598X，E-mail：zhangwq@llas.ac.cn。
*本文系中国科学院文献情报能力建设专项“中国科学院机构知识库功能扩展”项目(项目编号: Y5ZG08100)的研究成果之一。

作者公开发表的论文，提供标准接口支持数据共享[8]。中国科学院机构知识库系统与 iSwitch 合作，各研究所 IR 自动收割并导入 iSwitch 分发的本机构知识产出数据。

基于 iSwitch 的 IR 内容建设主要是通过定时任务调度与 FTP 协议进行数据同步，对文件包、文件解析提取其中的分发批次信息、知识产出元数据并将其预加载到数据库，为管理员提供数据导入管理功能以半自动化方式实现数据的最终导入。此外，还提供了出错管理、已导入数据批量更新、审计等功能。

2 功能框架

iSwitch 使用 FTP 通信协议实现数据共享。每当接收到出版社新推送的数据时，则解析识别作者的机构、资助机构，再把知识产出数据分发到相应机构的文件目录下。基于开放获取论文推送转发服务系统 iSwitch 的 IR 内容建设主要功能包括数据同步、批次数据浏览与导入、已导入数据管理、作品认领与审计等。系统总体功能框架设计如图 1 所示:

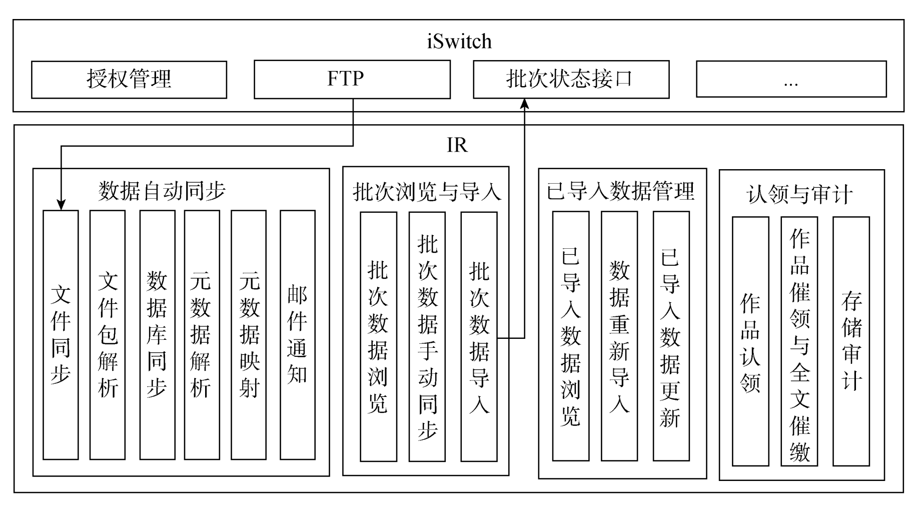


图 1 系统功能框架

IR 直接使用的 iSwitch 服务主要有授权管理、FTP 服务器、批次状态接口。

(1) 授权管理: iSwitch 采用基于 IP 的访问授权控制。一般由 IR 开发人员直接登录 iSwitch 授权系统完成相关信息的注册。

(2) 数据自动同步: IR 通过定时任务调度从 iSwitch 的 FTP 服务器同步本机构知识产出，数据下载到本地缓存后，通过文件解析、元数据解析与映射转换，保存到数据库并给 IR 管理员发送导入提醒邮件通知。

(3) 批次浏览与导入: 支持 IR 管理员浏览批次列表以及批次下的知识产出详细信息。由于自动同步存在网络传输中断可能以及管理员实时同步数据的需求，系统支持对某些或全部批次列表进行手动同步。最终由 IR 管理员完成数据执行，导入时需确定导入方式、目标专题、导入字段等。某一批次数据全部导入后，向 iSwitch 的批次状态信息接口反馈导入状态。

(4) 已导入数据管理: 支持 IR 管理员按导入方式分类浏览已导入数据，支持对导入后又被删除数据的状态记录以及对出错数据或已删除数据的重新导入。考虑到 iSwitch 原始数据更新的可能，提供对已导入知识产出元数据自动更新的功能。

(5) 认领与审计: iSwitch 导入的数据一般只有元数据而不包含全文。IR 通过作品催领与全文催缴、存储审计等功能，审计并提醒用户及时完成作品认领、全文上传。

3 关键功能的设计与实现

3.1 数据同步

iSwitch 分发的知识产出数据在 FTP 服务器上的组织结构如"出版商→机构→批次→文章"。一个批次对应一次文件分发，一次分发一般包含多篇文章。每篇文章以 ZIP 文件包的形式存储，其中包括出版商原始版本的元数据描述文档(一般为 XML 格式)以及 iSwitch 使用 JATS(Journal Article Tag Suite)标准[9]重新编码的文档。

iSwitch 提供单独的批次及其知识产出列表描述

服务接口，但 FTP 存放数据的目录结构本身包含了这些信息，且原始数据的获取也需要从 FTP 服务器下载。所以，IR 通过直接读取 FTP 目录结构获取批次信息。为了方便文件同步，IR 在本地建立缓存目录并采取与 iSwitch 的 FTP 批次文件存储目录相一致的组织结构。

IR 创建了基于 Quartz 框架[10]的定时任务对 iSwitch 数据进行自动收割。目前，收割频率默认是每周一次。FTP 文件下载功能主要是基于 Apache Commons-Net 程序包的 FTP Client 模块提供的公共服务接口实现。从元数据描述 XML 文档中解析读取元数据使用了 JDOM 组件。批次数据自动同步程序处理流程如图 2 所示：

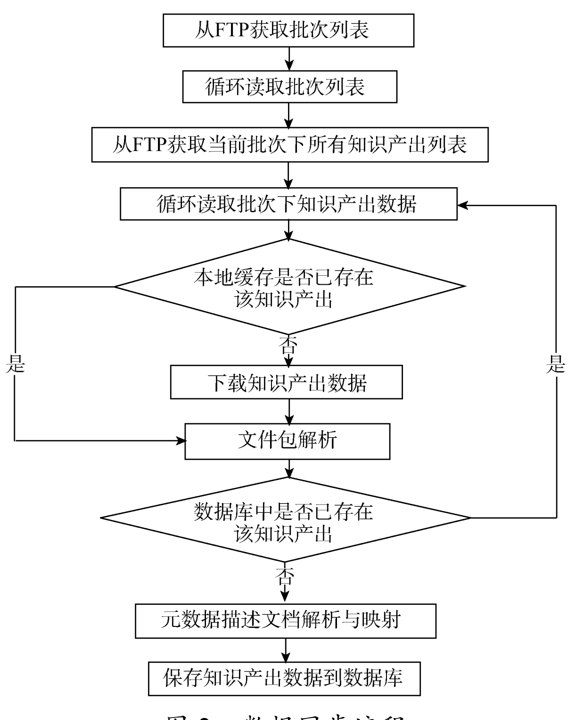


图 2　数据同步流程

同步任务启动后，IR 从参数配置中取得机构的正式名称，构造格式如"iSwitch 域名/机构名称"的 URL。访问该地址获取 iSwitch 分发给当前机构的所有批次目录，再循环读取每个批次下所有知识产出文件列表，检查每个知识产出在本地缓存是否存在，如果不存在则下载。下载到的每篇知识产出数据为 ZIP 压缩文件，文件名称中包含了分发年月与文章唯一编号信息，如"201410.00024"。IR 将 ZIP 文件名作为每篇文章的唯一标识。依据该标识查询数据库判断文章是否已加载到数据库。如果没有，IR 解压后提取元数据描述文档并解析元数据。考虑到数据重新编码有可能带来的元数据项丢失，IR 中选择直接读取出版社版本的描述文档。由于 IR 底层数据描述是基于 DC 元数据框架，解析得到的元数据需要做元数据映射处理之后存入 IR 数据库。不同出版社使用的编码格式一般不同，在处理不同来源出版社的数据时需要创建与其对应的数据解析器。同步过程中 IR 还会检测 iSwitch 中已删除，但本地已加载到数据库中的批次、知识产出数据，将这些数据删除(该操作不影响已导入数据)。自动同步完成后，如果此次同步下载到新数据，则会向 IR 管理员发送邮件通知提醒及时完成数据导入。

除了自动同步，IR 还提供手动同步功能。手动同步除了实现类似自动同步的功能外，还支持可选同步哪些批次，以及是否重新从元数据描述文档中解析元数据并更新到数据库。

3.2　数据导入

同步到 IR 尚未导入的知识产出数据与 IR 中已正式导入、可公开访问的数据在数据库中的存储组织方式基本一致，只是将其标记为未导入且不公开访问。IR 提供数据导入管理功能，由 IR 管理员实现对 iSwitch 数据的导入。导入操作之所以没有实现完全自动化，是因为 IR 中的知识产出是以专题为单元组织的，需要由管理员确定每篇文章属于哪个专题。此外，对于重复数据的处理也需要由管理员确定导入方式、导入元数据项等。

导入功能支持三种可选查重方式：题名、内容类型+题名、内容类型+(出版社文章编号或 DOI 或题名)。导入方式有三种：新增、续补、跳过。新增是无论 IR 中是否已存在都会创建新的记录；续补是针对已存在数据，使用 iSwitch 数据更新已有的知识产出元数据；不属于本机构的导入时可以选择跳过。

为了减少导入过程中管理员的人工操作，在加载待导入数据时，系统会提前做一些数据预处理，主要是对已存在数据的检测与目标专题的确定。通过默认的查重方式检测每条待导入数据是否已存在并在导入管理界面显示查重结果。对于 IR 中已存在的，目标专题默认与已存在数据所属专题保存一致，导入方式也选择续补。对于 IR 中不存在的，导入方式默认选择新增；IR 根据作者名称与系统用户别名库的模糊匹配，

预先判断该知识产出可能的所属用户，并将该用户发表期刊论文所属专题作为待导入数据的默认专题。IR 管理员在执行导入时可修改默认的专题和导入方式。

导入管理界面如图 3 所示:



图 3　iSwitch 数据导入管理

页面上方是针对当前导入的一些公共参数项设置，包括要导入的批次、目标内容类型、查重方式、每页显示数据量、导入哪些字段等。页面下方是当前批次未导入知识产出列表，列表左侧是通过解析 iSwitch 原始数据保存在数据库中的未导入条目，其中对中国科学院作者及其地址信息用黄色背景高亮显示以方便管理员导入时检查文章是否属于自己机构；列表右侧是通过所选查重方式检测到的系统已存在条目。每条数据下方是目标导题与导入方式选项。目前的功能支持一次导入多条数据，左右分栏的方式清晰明了地显示了未导入及其对应的已导入数据，从而最大程度地简化了 IR 管理员的操作，提高工作效率。

3.3　容错功能

基于 iSwitch 的 IR 内容建设，从 iSwitch 分发到 IR 的数据下载、解析、导入，甚至导入系统之后整个流程的每个环节都有可能出错。主要出错现象及 IR 的解决方案如下:

(1) iSwitch 分发出错。由于作者在不同期刊上发表文章时填写的作者机构不统一，造成了 iSwitch 在数据解析时可能将一些不属于本机构的数据分发过来。对于这种情况，IR 在导入时由管理员人工识别并选择“跳过”方式导入。

(2) 同步任务出错。同步任务时的出错主要表现在网络传输中断，以及 IR 对元数据描述文档中包含的特殊编码格式、特殊字符解析出错。这些出错会造成同步到 IR 的数据与 iSwitch 原始数据不一致。对此，IR 为管理员提供了手动同步功能，同步时还可选择是否重新加载 iSwitch 原始数据。当发现程序存在解析错误时，通过升级程序代码并重新执行手动同步任务来解决此问题。对于已经导入到系统中的数据，IR 也提供了批量更新功能，支持管理员选择更新特定元数据项。

(3) 已跳过、已删除的重新导入。有些以“跳过”方式导入的 iSwitch 数据有可能在后期发现确实是本机构的产出。此外，有些已导入数据会被管理员有意无意地删除，后期又想恢复这些数据。针对这种需求，IR 在数据成功导入后，并没有删除数据库中的 iSwitch 原始数据条目，而是将导入数据作为单独的条目保存，并保留两者之间的关联。这样，就可以记录已导入数据是否被删除；而且，iSwitch 原始数据有更新时，可以关联更新已导入数据。基于以上底层设计，IR 支持将已跳过、已删除的数据重新标记为未导入并重新导入。

3.4　审 计

IR 中的 iSwitch 数据审计功能，支持系统管理员实时了解共有多少批次、每个批次的导入情况，以及已导入作品的认领情况、全文上传情况等。对 iSwitch 来说，需要了解每个机构分发的数据是否已下载、某一批次数据是否已全部导入 IR 等。IR 在数据同步、导入过程中会及时向 iSwitch 返回批次数据的下载与导入情况以支持 iSwitch 数据审计需求。

(1) 研究所 IR 的 iSwitch 数据审计。研究所 IR 层面的审计是在 IR 内部实现的。IR 会记录每个批次知识产出的导入状态与导入方式，从而支持对批次导入进度以及按导入方式对已导入数据做分类统计。IR 的作品催领与全文催缴功能会实时保存用户与作品间的关联关系、作品认领及作品的全文存储状态，支持管理员对未认领作品及已认领但未提交全文作品的审计，并对相关用户批量发送任务邮件通知。

(2) 对 iSwitch 审计的支持。iSwitch 关于 IR 数据使用情况的审计，其中需要与 IR 系统协作完成的部分目前主要是对批次是否已成功导入状态的获取。IR 在

管理员执行批次导入操作时，会检查当前批次是否已导入完成。如果全部导入成功，则向 iSwitch 提供的接口返回批次唯一标识与状态信息。

4 应用效果评估

目前，iSwitch 已完成对 WoS(Web of Science)收录的中国科学院作者产出文章的历史回溯并支持对最新数据的自动接收与分发。截至 2015 年 12 月 8 日，中国科学院已有 83 家研究所的 IR 部署了 iSwitch 数据监测导入功能。根据中国科学院网格系统(IR Grid)[11]对研究所 iSwitch 数据下载导入情况的统计，已成功导入的 iSwitch 来源知识产出数据有 67 024 条，有全文的共 49 885 条，全文存储率达 74%以上。由于 iSwitch 不断接收并分发新的知识产出，该统计数字会随着时间不断增长。IR 导入 iSwitch 数据最多的 10 个研究所如表 1 所示：

表 1 iSwitch 数据导入 TOP10 研究所

研究所名称	数据导入量
中国科学院大连化学物理研究所	10 692
中国科学院过程工程研究所	4 264
中国科学院海洋研究所	4 189
中国科学院昆明植物研究所	3 969
中国科学院武汉物理与数学研究所	2 181
中国科学院心理研究所	2 102
中国科学院西安光学精密机械研究所	1 725
中国科学院水生生物研究所	1 396
中国科学院化学研究所	1 348
中国科学院水利部成都山地灾害与环境研究所	917

WoS 提供通过收录文章内部 ID(WoS 入藏号)获取其被引信息(包括被引数、文章在 WoS 中的链接、引用文章链接、相关文章链接等)的接口[12]，IR 中历史 WoS 数据一般不包含该元数据。使用 iSwitch 分发数据对已有数据补录后，可以在浏览知识资源时实时显示文章在 WoS 中的被引相关数据。

5 结 语

iSwitch 从各出版商获取数据、按机构分发，IR 下载并导入 iSwitch 已分发数据。通过两个系统之间的协作，实现了已公开发表论文从出版社到作者机构 IR 的自动推送。此前，这些数据都需要人工整理提交到 IR。基于 iSwitch 的 IR 内容建设，是科研人员参与积极性不高背景下的一种较为理想的 IR 内容建设模式。这种方式不仅减轻了科研人员、IR 管理员等相关参与者的负担，也避免了人工操作过程中可能造成的元数据出错或丢失问题。

系统在使用过程中发现还存在一些不足。例如，有些研究所的反馈，无法从 iSwitch 获取到 WoS 最新收录的数据，iSwitch 对新数据的分发存在滞后性。还存在 WoS 中有收录的数据，但 iSwitch 没有分发的现象。以上问题与出版社本身提供数据的完整性、推送频率以及 iSwitch 自身的数据分发机制有关。此外，由于作者机构填写不规范、不同来源期刊的作者机构不统一、机构历史名称变化等因素造成有些文章无法正确地分发到其真正的所属机构。IR 在数据导入自动化方面有待进一步提高。例如，现在对于已存在数据仍需要系统管理员确认导入，后期可以让管理员预先定义对已存在知识产出的处理规则。在数据同步时，如果 iSwitch 数据与 IR 已导入数据的题名及第一作者相同，根据预定义规则直接导入。

iSwitch 暂时只支持对 WoS 来源数据的分发，希望以后可以支持更多出版社开放获取论文的自动获取与转发。两系统在实际运行过程中，不断改进优化，使系统间的交互更加顺畅、功能更趋完善、自动化程度更高，让基于 iSwitch 的 IR 内容建设切实解决存缴难的问题。

参考文献：

[1] How to Increase Content in OA Repositories — What Can Be Learnt from the Special Case of the Research Project PEER-Publishing and the Ecology of European Research [EB/OL]. [2015-12-10]. http://www.peerproject.eu/fileadmin/media/ppt_about_peer/PEER-How_to_increase_content_in_repositories_April2012.pdf.

[2] 张晓林，梁娜，钱力，等. 开放获取论文推送转发服务系统 iSwitch：概念、功能与基本框架[J]. 现代图书情报技术，2014(10): 4-8. (Zhang Xiaolin, Liang Na, Qian Li, et al. Router Service Engine iSwitch for Open Access Articles: The Concept, Strategy, and Framework [J]. New Technology of Library and Information Service, 2014(10): 4-8.)

[3] Duranceau E F, Rodgers R. Automated IR Deposit via the SWORD Protocol: An MIT/BioMed. Central Experiment

[J/OL]. UKSG Series, 2010, 23(3): 212-214. http://dx.doi.org/10.1629/23212.
[4] Italian National Institute of Geophysics and Volcanology [EB/OL]. [2015-12-10]. http://www.ingv.it/en/.
[5] Lewis S, De Castro P, Jones R. SWORD: Facilitating Deposit Scenarios [J/OL]. D-Lib Magazine, 2012, 18(1-2). http://www.dlib.org/dlib/january12/lewis/01lewis.html.
[6] Jisc Publications Router [EB/OL]. [2015-12-10]. http://broker.edina.ac.uk/information.html.
[7] 中国科研论文数据交换中心[EB/OL]. [2015-12-10]. http://iswitch.las.ac.cn/. (Router Service Engine iSwitch for Open Access Articles [EB/OL]. [2015-12-10]. http://iswitch.las.ac.cn/.)
[8] 师洪波, 钱力, 张晓林, 等. 开放获取论文推送转发服务系统 iSwitch: 论文接收与解析[J]. 现代图书情报技术, 2015(6): 1-6. (Shi Hongbo, Qian Li, Zhang Xiaolin, et al. Router Service Engine iSwitch for Open Access Articles: Articles Reception and Resolving [J]. New Technology of Library and Information Service, 2015(6): 1-6)
[9] Journal Article Tag Suite [EB/OL]. [2015-12-10]. http://jats.nlm.nih.gov/.
[10] Quartz Scheduler [EB/OL]. [2015-12-10]. https://quartz-scheduler.org/.
[11] IR Grid [EB/OL]. [2015-08-14]. http://www.irgrid.ac.cn.
[12] Article Match Retrieval [EB/OL]. [2015-08-14]. http://wokinfo.com/products_tools/products/related/amr/.

作者贡献声明:

张旺强: 论文起草, 系统功能的详细设计与实现;
祝忠明: 提出研究思路, 论文最终版本修订;
姚晓娜: 统计各研究所 IR 下载导入 iSwitch 分发论文数据;
刘巍: 系统功能的升级部署。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。
[1] 师洪波. 201410.00016.zip. iSwitch分发的单篇文章样本数据.

收稿日期: 2015-12-14
收修改稿日期: 2015-12-29

Building Institutional Repository with iSwitch Service

Zhang Wangqiang　Zhu Zhongming　Yao Xiaona　Liu Wei
(Lanzhou Library, Chinese Academy of Sciences, Lanzhou 730000, China)

Abstract: [**Objective**] This study aims to help an organization automatically download its empolyees' open access papers from iSwitch, and then import these articles to the institutional repository. [**Methods**] We first synchronized data from iSwitch through timing task scheduling based on FTP protocol. Second, we parsed files and saved metadata to the database in advance. Some functions, such as import process and data management, as well as audit, were also provided. [**Results**] Papers could be automatically synchronized from iSwitch and then imported to the institutional repository by the system administrator. We have successfully analyzed and imported more than 60, 000 items from Web of Science. [**Limitations**] The accuracy and timeliness of the service by iSwitch need to be improved. The data import function of the institutional repository should also be optimized for better services. [**Conclusions**] The high quality institutional repositories built on iSwitch, which significantly relieve burden of researchers and system administors, should be promoted.

Keywords: Open Access　Institutional Repository　iSwitch　Content recruitment